

Classifying informative and imaginative prose using complex networks

Henrique F. de Arruda

*Institute of Mathematics and Computer Science
University of São Paulo, São Carlos, São Paulo, Brazil*

Luciano da F. Costa

*São Carlos Institute of Physics
University of São Paulo, São Carlos, São Paulo, Brazil*

Diego R. Amancio*

*Institute of Mathematics and Computer Science
University of São Paulo, São Carlos, São Paulo, Brazil*

Statistical methods have been widely employed in recent years to grasp many language properties. The application of such techniques have allowed an improvement of several linguistic applications, which encompasses machine translation, automatic summarization and document classification. In the latter, many approaches have emphasized the semantical content of texts, as it is the case of bag-of-word language models. This approach has certainly yielded reasonable performance. However, some potential features such as the structural organization of texts have been used only on a few studies. In this context, we probe how features derived from textual structure analysis can be effectively employed in a classification task. More specifically, we performed a supervised classification aiming at discriminating informative from imaginative documents. Using a networked model that describes the local topological/dynamical properties of function words, we achieved an accuracy rate of up to 95%, which is much higher than similar networked approaches. A systematic analysis of feature relevance revealed that symmetry and accessibility measurements are among the most prominent network measurements. Our results suggest that these measurements could be used in related language applications, as they play a complementary role in characterizing texts.

I. INTRODUCTION

The ever-growing amount of available documents in the Web has propelled the development of statistical natural language processing methods in recent years. Examples of related applications trying to “understand” unstructured data include machine translation [1], text summarization [2–5], information retrieval [6, 7] and content analysis [8]. An application of special importance for the organization of electronic data is the classification task, which automatically assigns one or more labels for a word, sentence, paragraph or entire documents [9–12]. Traditional textual categorization methods usually serve to identify the relevance of texts (i.e. whether it is a spam or not) or the meaning conveyed by words and expressions [13–17]. Recent classification tasks, however, have emphasized other textual aspects. For example, the categorization of texts according to their polarity (e.g. positive or negative) has become a relevant task for analyzing e.g. customer reviews [18] or global variations in mood via polarity analysis of twitter messages [19]. Note that most of these classification tasks are dependent on text content, since the presence of one or more specific words provides clues about the classes being inferred. While the semantic content is crucial for the success of these applications, the structure of texts might play an important role in classifications problems where the semantics of words is not crucial for the purpose of categorization. This is the case of identifying the style of

texts, since documents on the same subject might display different writing styles. In contrast to semantic-based traditional classification tasks, in this paper we probe the relevance of textual structure to provide useful features for text classification. More specifically, we probe how textual structure depends on two distinct stylistic writing styles: imaginative and informative prose. The structure and organization of texts is studied via networked models, an well known representation of complex systems.

Networks are discrete models that basically represent the interrelations between interacting agents in a complex system. Owing to the simplicity and generality of the model, it has been employed to model a myriad of real and artificial systems [20]. Despite being very different in nature, networks modelling distinct systems share several structural patterns [21]. Of special relevance to this paper, are the networked models of language and texts, which have been useful to unveil universal properties including the scale-free and small-world phenomena [22]. In practical terms, networked models have been useful to grasp several features of texts, such as quality [23], complexity [24] and authenticity [25]. Particularly, in this study, we used the so-called word adjacency model, which is a approximation of text networks formed by syntactical links [26]. Because the topological analysis of word adjacency networks does not depend on the interpretation of texts, it has been applied with relative success to study the underlying structure of texts, even

when textual content is entirely unknown [25]. Here we applied such representation to discriminate informative and imaginative prose. We have extended the traditional model in a twofold fashion: (i) we analyzed the local structure of particular nodes (words); and (ii) we considered novel topological/dynamical that are able to unfold the general structure of the network of concepts. As we shall show, both extensions provided competitive classification performance when compared to traditional stylometry methodologies. In special, we have found that symmetries and accessibilities were the most important network measurements. We believe that the proposed extended model could be used to improve the performance of several related problems where the textual structure plays a prominent role in the characterization of documents [27].

This paper is organized as follows. In Section II, we present the word adjacency model. Section III describes the pattern recognition methods used to perform the classification. This section also presents a method for measuring feature relevance in a multivariate fashion. The results obtained with the proposed methodology are described in Section IV. Finally, Section V provides a perspective for further studies and improvements of the model.

II. REPRESENTATION AND CHARACTERIZATION OF TEXTS AS NETWORKS

In this section, we present how a text can be represented as a complex network. We also swiftly describe the main topological network measurements employed to characterize the structure of networks.

Modeling texts as complex networks

A complex network can be represented as a graph, which is defined as a set of nodes and edges [21]. An usual representation of a network is the adjacency matrix $A = \{a_{ij}\}$. The elements are defined as:

$$a_{ij} = \begin{cases} 1 & \text{if there is a link } i \rightarrow j, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

In the text mining community, several networked representations of texts have been proposed [28]. If the stylistic properties of texts are relevant for the task being tackled, syntactical relations are employed to establish the links between words [26]. If the application requires the extraction of semantic features, words are connected according to semantic relations, such as those present e.g. in the WordNet [29] or in free association graphs [30]. In this study, we aim at grasping textual features that

are independent of semantic content. For this reason, we used a model that is able to capture stylistic textual features [12, 31]. This model, henceforth referred to as word adjacency model, denotes each distinct word as a node. The edges are established between words appearing in the same context. It has been shown that if one considers the context as the interval of two adjacent words, most of the syntactical relations are recovered [26]. This model has been successfully applied to study many language applications and related systems [31].

To create a word adjacency network, usually some pre-processing steps are applied. First, all punctuation marks, line breaks, spaces, numbers and special characters are removed. Particularly, we are mostly interested in the relationship between words conveying semantic information. For this reason, *stopwords* (or function words) can be optionally removed from the analysis. In the next step, a lemmatization step is performed in order to map words representing the same concept into the same node. To assist the lemmatization process, the part-of-speech tag of each word is extracted according to the procedure described in [32]. The part-of-speech labelling is required to solve ambiguities, because the same word form might be mapped into distinct lemmas. After the pre-processing step, each remaining distinct word becomes a node and edges are established between adjacent words.

Complex network measurements

Currently, there are more than a hundred measurements employed to characterize the topological structure of networks [33]. Some measurements might depend not only on the structure, but also on a dynamical process (e.g. random walks) occurring on the structure. Below we swiftly describe the measurements used in this study.

- **Number of nodes (V):** in a word adjacency network, the number of nodes is the set of different words in the text. In other words, the number of nodes is the vocabulary size of the pre-processed text.
- **Degree (k):** the simplest connectivity measurement is the node degree [34], which corresponds to the total number of edges connected with node i . This measurement is defined for directed networks as $k^{(\text{in})} = \sum_i a_{ij}$ and $k^{(\text{out})} = \sum_j a_{ij}$ for in- and out-degree, respectively. If one considers the undirected and unweighted version of the network, the degree $k_i = \sum_j a_{ji} = \sum_j a_{ij}$ can be understood as the number of distinct bi-grams that a given word appears. If one considers edges weights, then the degree is proportional to the word frequency.
- **Neighborhood connectivity (N):** this measurement is defined as the number of nodes that can

be reached when, starting from the reference node, walks of length h are performed. Note that the traditional degree measurement is recovered when $h = 1$.

- **Clustering coefficient (cc):** given a node i , the probability of its neighbors to be connected is called clustering coefficient (cc_i) [35]. This measurement is defined as

$$cc_i = \frac{3N_{\Delta}(i)}{N_3(i)}, \quad (2)$$

where $N_{\Delta}(i)$ is the total number of triangles (i.e. a click comprising three nodes) connected with node i , and $N_3(i)$ denotes the number of connected triples, which is defined as the amount of different connections between i and each pair of nodes. This measurement is traditionally used to quantify the local connectivity of real-world networks [36]. In word adjacency networks, this measurement has been applied to quantify the specificity of words according to the number of distinct contexts in which they appear [37].

- **Betweenness centrality (B):** to define this measurement, consider all paths connecting any pair of nodes in the network are followed via shortest paths [38]. The betweenness of a node u is defined as being proportional to the number of paths that passes through node u . More specifically,

$$B_u = \sum_{ij} \frac{\sigma(i, u, j)}{\sigma(i, j)}, \quad (3)$$

where $\sigma(i, u, j)$ is the number of shortest paths between i and j that passes through node u and $\sigma(i, j)$ is the total amount of shortest paths between i and j . According to equation 3, the betweenness centrality can be interpreted as the network flow [39–41], which is a relevant quantity for the analysis of robustness of power-grid networks [42, 43]. When applied to the analysis of text networks, this measurement has been interpreted as being useful to quantify the generality of words in which the word appears [44], which is in part motivated by the use of this measurement in community detection methods [45]. Unlike the clustering coefficient, the betweenness centrality uses the global connectivity information to quantify the specificity/generality of concepts [44].

- **Closeness centrality (C):** unlike the betweenness centrality, which is based on the *number* of shortest paths, the closeness centrality [46] uses the *length* of the shortest paths. If d_{ij} is the shortest distance between nodes i and j , the closeness centrality is calculated as $C_i = V^{-1} \sum_j d_{ij}$. Geodesic paths have

been reinterpreted in the context of text networks as a measure of word relevance. Actually, a word is deemed relevant if it is very frequent in the text or if it appears related to other very frequent words [37].

- **Eccentricity (E):** this measurement quantifies the maximum geodesic distance between the reference node and all other nodes [47]. Therefore, the maximum eccentricity value corresponds to the network diameter. This measurement is calculated for each node i as $E_i = \max_j(d_{ij})$.
- **Eigenvector centrality (Ec):** the eigenvector centrality can be understood as an extension of degree centrality [48], because the relevance of the reference node relies both on the *number* and *relevance* of neighbors. Considering the adjacency matrix A , the eigenvector centrality is defined as the eigenvector associated with the leading eigenvalue. There are many linguistic applications that uses this centrality measurement. It has been applied, for example, in the text summarization task in order to select the most relevant extracts in texts modelled as graphs [4].
- **PageRank (Pr):** the PageRank is widely known to be part of the Google’s web search [21, 49]. In texts networks, this measurement has been successfully applied e.g. to disambiguate word senses [50]. This measure is based on the eigenvector centrality, and it is defined in matrix terms as

$$Pr = \alpha AD^{-1}Pr + \beta 1, \quad (4)$$

where α and β are positive constants (conventionally $\beta = 1$), 1 is a vector $(1, 1, 1, \dots)^T$ and D is a diagonal matrix represented as

$$D_{ij} = \begin{cases} \max\{k_i^{(\text{out})}, 1\} & \text{if } i \neq j, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

In contrast with eigenvector centrality, PageRank considers a weighted sum of neighbors importance reflecting the neighbors degree. In this way, the relevance associated to a node is proportionally transferred to its neighbors.

- **Accessibility ($A^{(h)}$):** the accessibility is an extension of the concept of neighborhood connectivity because it measures the effective number of nodes reached at the h -th concentric level [51]. The effective number of nodes accessed after h steps is computed considering the distribution of probabilities of access via self-avoiding random walks. Mathematically, it is defined using the Shannon entropy [52] of the probabilities of access at the h -th

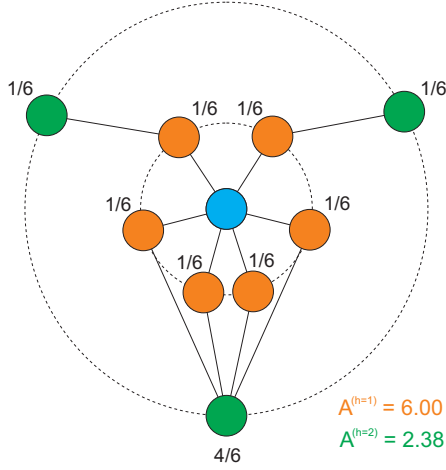


FIG. 1. Quantification of accessibility for the local neighbourhood of the blue node. The probabilities near nodes represent the probabilities of access in self-avoiding random walks of length $h = 1$ (orange nodes) or $h = 2$ (green nodes). In the second hierarchical level, the accessibility is lower than the total number of nodes (green nodes). This occurs because the access to the second level is uneven.

concentric level:

$$A_i^{(h)} = \exp \left(- \sum_j p_{ij}^{(h)} \log p_{ij}^{(h)} \right), \quad (6)$$

where $p_{ij}^{(h)}$ is the probability of a walker starting from i to reach node j in h steps. In text networks, this measurement has been applied to generate summaries and to identify keywords and styles [2]. An example of the computation of accessibility is shown in Fig. 1.

- **Generalized accessibility (Ag):** the generalized accessibility is an extension of the accessibility that does not rely on a particular length of walk. Instead, the probabilities of transition are computed considering all possible lengths, which is implemented via definition of a modified random walk, the so called accessibility random walk [53]. This measurement is defined as

$$Ag_i = \exp \left(- \sum_j \mathbf{P}_{ij} \log \mathbf{P}_{ij} \right), \quad (7)$$

where \mathbf{P} is a quantity that depends on the probability of the transition $i \rightarrow j$ considering walks of variable length. The matrix \mathbf{P} representing the probability of transition is calculated as $\mathbf{P} = \mathbf{W}/e$, where

$$\mathbf{W} = \sum_{k=0}^{\infty} \frac{1}{k!} \mathbf{P}^k = e^{\mathbf{P}}. \quad (8)$$

Note that, according to equation (8), the highest weights are assigned to the nearest nodes. The generalized accessibility has been applied e.g. to identify influential spreaders in spatial networks [53].

- **Symmetry (S):** the symmetry concept is found in many real systems [54, 55]. Symmetric properties can also occur in written texts as a consequence of grammatical or stylistic constraints [59]. For this reason, we have quantified this property in word adjacency networks. To model such property, recently, some network measurements have been created [56–59]. In this paper, we used the quantities introduced in [58, 59], as it allows to capture symmetric patterns in a multi-scale fashion. The definition of symmetry measures rely upon the characterisation of *hierarchic levels*. The hierarchic level $\Gamma_h(i)$ for a given node i is the set comprising all nodes h hops away from i . The symmetry measures are based on the accessibility measurement, because the same network dynamics is taken for the analysis. In addition, the symmetry measurement can be seen as a normalization of the accessibility. Thus, using self-avoiding random walks, a node is considered to be symmetric if the access to its neighbors (in a given hierarchic level) is symmetric. The symmetry (or regularity) of the access is measured in terms of the entropy:

$$S_h(i) = \frac{\exp \left\{ - \sum_{j \in \Gamma_h(i)} p_{ij}^{(h)} \log p_{ij}^{(h)} \right\}}{|\Gamma_h(i)| + \sum_{r=0}^{h-1} \eta_r}, \quad (9)$$

where η_r denotes the total number of dead ends in the r -th hierarchical level and $p_{ij}^{(h)}$ is the same quantity used to define the accessibility in equation (6). There are two variations of the quantity proposed in equation (9). The backbone symmetry (Sb), a variation of the the concept of radial symmetry, removes all edges between nodes in the same hierarchical level. The merged symmetry (Sm), on the other hand, is based on the concept of angular symmetry, which can be obtained by merging linked nodes in the same hierarchical level. To exemplify both variations of the symmetry concept, we show in Fig. 2 the transformations applied to a hierarchical neighborhood before the computation of equation (9).

- **Modularity (Q):** a community structure is defined as a network subgraph with a large number of intra-links and a few edges connected to the others nodes of the network. To quantify whether a network is organized in communities, the modularity compares the number of internal links with the expected value of the same quantity in a equivalent

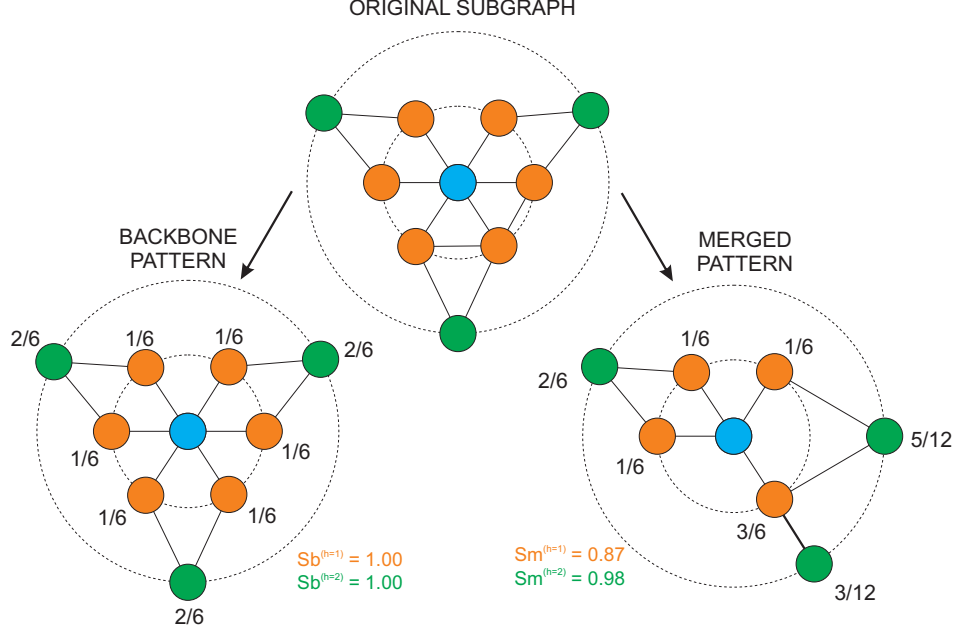


FIG. 2. Example of merged and backbone patterns for the quantification of local symmetry. The fractions represent the probabilities to reach a given node in a random walk of length $h = 1$ (orange) or $h = 2$ (green). To create the backbone pattern, edges among nodes in the same hierarchical levels are removed. Differently, merged patterns are computed by merging connected nodes into a single node.

random network [60]. This quantity is computed as

$$Q = \frac{1}{2M} \sum_i \sum_j \left(a_{ij} - \frac{k_i k_j}{2M} \right) \delta(c_i, c_j), \quad (10)$$

where $M = 1/2 \sum_{ij} a_{ij}$ is the total number of edges in the network, c_i and c_j are the communities to which nodes i and j belong, and

$$\delta(c_i, c_j) = \begin{cases} 1 & \text{if } c_i = c_j, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

Usually, word adjacency networks display low values of modularity. A more consistent organization in communities can be found e.g. in semantic networks such as the WordNet [29].

Characterization of texts with complex networks

So far we have presented several topological/dynamical measurements of complex networks. The objective here is to use these quantities to characterise styles in texts. Note that several measurements are locally defined, i.e. each node possess a value. There are several possibilities to use these local measurements to characterise the networks. In this paper, we have used the following three distinct methodologies:

- **Global strategy without stopwords (GS):** in this approach, we sum up the local measurements to characterise the networks. The most natural summarisation procedure is to take the average $\langle X \rangle$, where $\langle X \rangle = V^{-1} \sum_i X_i$ and X is a local measurement. We also used the following quantities: the standard deviation $\sigma(X)$, the median (\tilde{X}), the maximum value ($\max(X)$) and the minimum value ($\min(X)$). The only global measurement, the modularity, was also used considered in this strategy. Following several approaches for grasping textual features with networks, we have removed all stopwords before the formation of the networks. These words were disregarded from the analysis because they just serve to connect content words in the word adjacency model.
- **Local strategy without stopwords (LS):** in this approach, the value of each measure X for each word becomes a feature. Similarly to the GSW, all stopwords are removed from the analysis. Because features are defined for each word, global networks measurements are not considered in this case.
- **Local strategy with stopwords (LSS):** this is the same local approach adopted in the LS method. However, this variation also considers all stopwords in the analysis.

III. PATTERN RECOGNITION AND EVALUATION

In this section, we present the methodology for analyzing the relationship between the texts and the categories (informative and imaginative). More specifically, we describe the pattern recognition methods employed and the methods to compute the quality of the classification and relevance of the proposed features (see Section II).

Pattern recognition methods

To study the relationship between complex network measurements and text style, we used a feature selection algorithm and three different supervised classifiers. The method used to select the features was the information gain [61], which is a supervised attribute filter known as mutual information [62]. Given the random variables X and Y , the mutual information $I(X, Y)$ is computed as

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (12)$$

where $p(x)$ and $p(y)$ are probability functions and $p(x, y)$ is the joint probability.

The information gain corresponds to the mutual information when X is the values obtained for a given attribute and Y is a vector of corresponding classes. This technique is used to create a decreasing sorted ranking of relevance. Thus, the most relevant attributes, i.e. the ones with the highest values of information gain, are selected to perform the classification. An important characteristic of this method is that the attributes are evaluated separately, i.e. the information of a given attribute does not influence the others.

In our experiments, the following pattern recognition methods were used:

- **Nearest neighbors:** the K nearest neighbors classifier (KNN) considers the local neighborhood of the test instance [63]. Given a test instance, the class chosen is the majority class in the set of the K nearest neighbors in the training dataset. Further details concerning this method can be found in [64].
- **Classification and regression tree:** this method represents the patterns found in the dataset as a tree, a data structure storing nested rules. Even though there are several tree algorithms, we chosen to use the classification and regression tree (CART) method [65] because it has some advantages as it is relatively simple for interpret and the predictor variables are not previously assumed [66]. A major advantage of tree-based pattern recognition algorithms the patterns found in the dataset are

not hidden from the user, as it happens in artificial neural networks methods [64].

- **Naive Bayes:** the Naive Bayes algorithm is based on the Bayes theorem [67]. Assuming feature independence, the correct class \tilde{c} of an instance is given by

$$\tilde{c} = \arg \max_{c_k} [\log P(c_k) + \sum_{f_j \in F} \log P(f_j | c_k)],$$

where c_k is one of the possible classes, $f_j \in F$ is a particular feature. To compute the quantity $P(f_j | c_k)$ we assumed that the likelihood of the features follows a bell shape [68].

We have chosen the aforementioned methods because they yield good performance when set with default parameters [69]. The evaluation of the performance of the methods when set with default parameters was performed with the “leave one out” algorithm [70]. This evaluation procedure consists in selecting one element of the dataset to be used as an test instance, while the remaining instances are used in the training phase. This procedure is then repeated until all instances of the dataset have been chosen as an test instance.

Quantifying feature relevance

To quantify the relevance of features for the classification task, the following method was used. Let $F = \{f_1, f_2, \dots\}$ be the set of attributes comprising Φ distinct attributes. We generate a set F_c comprising all 2^Φ combinations of features. For a particular classification method, we compute the accuracy rate obtained for each classifier in F_c . The accuracy rate is then employed to sort (in decreasing order) the classifiers in F_c . A function is associated for each attribute in F :

$$\Omega(f_i, k) = \sum_{j=1}^k \omega(f_i, j) \quad (13)$$

where $\omega(f_i, j) = 1$ if the j -th best classifier in F_c used the i -th attribute. If the i -th feature was not used in the j -th best classifier, then $\omega(f_i, j) = 0$. Note that the function Ω quantifies how frequent a given feature is in the best classifiers. If this function has a fast growth for low values of k , then it is a relevant feature because it appears in the classifiers with the highest accuracy rates. To quantify how frequent a given feature f_i is among the best classifiers, the following index of feature relevance can be defined:

$$R(f_i) = \sum_{k=1}^{2^\Phi-1} \Omega(f_i, k) = \sum_{k=1}^{2^\Phi-1} \sum_{j=1}^k \omega(f_i, j). \quad (14)$$

Unlike traditional index devised to measure the relevance of attributes, the index defined in eq. (14) takes into account the non-trivial inter-relationship between features [71].

IV. RESULTS AND DISCUSSION

In this section, we analyze the proposed technique for discriminating informative and imaginative prose. We also compare the proposed technique with other traditional natural language processing methods. In our experiments, we used the Brown University Standard Corpus of Present-Day American English (a.k.a. Brown Corpus) [72]. Because the set of informative texts comprises several short texts, for this class we have selected only the 126 longest texts. As such, in our experiments, each class is represented by the same number of instances. Each class can also be classified in subclasses. The set of informative texts used in this study comprises 80 scientific manuscripts, 30 miscellaneous texts and 16 biographies and related subjects. The set of imaginative documents comprises general fiction, romances, love stories and others. Note that we have not used this fine-grained description in our experiments.

Complex network approach

Following the steps in the methodology, we created a word adjacency network for each document in the dataset. The topological measurements were extracted and the 15 most relevant features were selected according to the information gain criterion. In the *global strategy*, the following features have been selected:

Vocabulary size: V ;
Degree connectivity: $\langle k \rangle$;
PageRank: $\bar{P}r$, $\langle Pr \rangle$, $\sigma(Pr)$ and $\max(Pr)$;
Clustering coefficient: $\sigma(cc)$ and $\langle cc \rangle$;
Closeness centrality: $\min(C)$, $\sigma(C)$, $\langle C \rangle$ and \tilde{C} ;
Generalized accessibility: $\langle Ag \rangle$; and
Betweenness centrality: $\langle B \rangle$ and \tilde{B} .

In this case, the accuracy rate reached a maximum value of 78% with the Naive Bayes algorithm. Note that this result is statistically significant, as the p -value associated with this accuracy rate is $p < 1.0 \times 10^{-10}$. The accuracy rate obtained for the other classifiers are shown in the first row of Table I.

When the *local strategy* (LS) was used to perform the classification, the accuracy rate improved by a large margin: all three classifiers reached an accuracy rate of 92%. The largest improvement in performance occurred for the KNN classifier; the accuracy rate went from 72% to 92%. The features employed in this case were:

TABLE I. Accuracy rate obtained with the three proposed network approaches. Note that, the most accurate results occur when the

Complex network approach	KNN	CART	Bayes
Global strategy without stopwords	72%	78%	75%
Local strategy without stopwords	92%	92%	92%
Local strategy with stopwords	95%	95%	95%

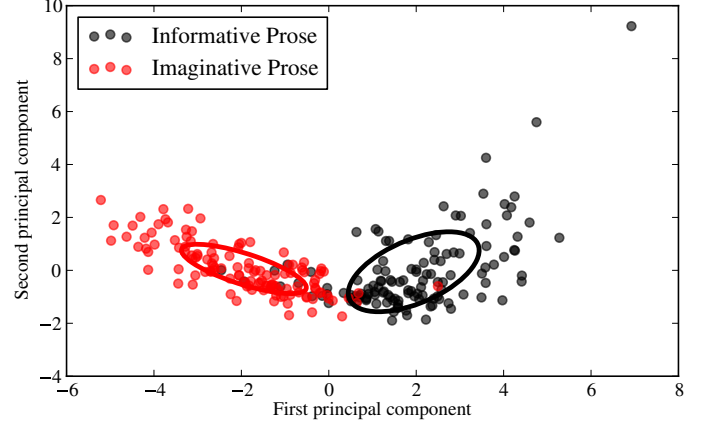


FIG. 3. Principal component analysis performed using local topological properties in networks formed with stopwords. To create the visualization, only the features with no correlation with the frequency were used.

Backbone symmetry: $Sb^{(h)}$, for $h = \{2, 3, 4\}$;
Merged symmetry: $Sm^{(h)}$, for $h = \{2, 3, 4\}$; and
Accessibility: $A^{(h)}$, for $h = \{2, 3\}$.

Note that these local measurements were chosen because they do not correlate with the frequency. The *local strategy with stopwords* (LSS) displayed a slight better classification performance. In this case, the accuracy rate reached 95% with KNN, CART, and Naive Bayes classifiers. The principal component analysis projection provided in Fig. 3 confirms the suitability of this network model for discriminating informative from imaginative prose. In this case, the features employed were:

Backbone symmetry: $Sb^{(h)}$, for $h = \{2, 3, 4\}$;
Merged symmetry: $Sm^{(h)}$, for $h = \{2, 3, 4\}$;
Accessibility: $A^{(h)}$, for $h = 2$; and
Generalized accessibility: (Ag) .

Note that, in both *local strategies*, the accuracy rates in the classification are much higher than the ones obtained with the *global strategy*, which suggests that a few words account for the informativeness of the topological approach. To better understand the factors

behind the network ability to discriminate informative from imaginative prose, we evaluated the relative importance of features employed in the best approach, i.e. the *local strategy with stopwords*. The method employed to quantify the relevance of features is described in the methodology. According to this method, the most relevant features, in decreasing order of relevance were:

- (i) Merged symmetry: $Sm^{(h=2)}(the)$
- (ii) Merged symmetry: $Sm^{(h=3)}(by)$
- (iii) Backbone symmetry: $Sb^{(h=4)}(by)$
- (iv) Merged symmetry: $Sm^{(h=3)}(an)$
- (v) Generalized accessibility: $Ag(have)$
- (vi) Merged symmetry: $Sm^{(h=4)}(by)$
- (vii) Generalized accessibility: $Ag(it)$
- (viii) Generalized accessibility: $Ag(by)$.

Note that the most relevant features are those related to the symmetry of specific words. Interestingly, this is consistency with recent results showing that symmetry measurements tend to be more discriminative than other traditional network measurements [73].

It is relevant to highlight that most of the network approaches for text classification focus on the global properties of networks. Our results reveal, conversely, that the informativeness of the topological strategy concentrates in a few nodes. Particularly, the informativeness was found to be mostly hidden in the symmetry patterns of specific function words. For this reason, we believe that the *local strategies* (LS and LSS) could be useful not only for the studied task, but also in several related tasks, where the topology of specific words plays a prominent role in characterizing texts.

Comparison with traditional methods

To compare the performance of the proposed technique with other traditional techniques, we first analysed if the classes can be discriminated via Latent Semantic Analysis [74], which considers as features the frequency of words. The projection obtained with this technique is shown in Fig. 4. Note that a good discrimination was obtained in this case, mainly because some words are more common in informative documents (e.g. *state*, *system* and *program*, while others occur more often in imaginative texts (e.g. *say* and *mr.*). A more accurate classification system based on stylistic attributes can be created if one considers as features the frequency of the most informative *stopwords*. To select the most informative stopwords, we used the information gain criterion. Using the *KNN* classifier (the best classifier), the performance reached 97% of accuracy. This high accuracy level can be observed in the principal component analysis provided in Fig. 5. Another traditional strategy in stylometry consists in counting the frequency of character bigrams [27].

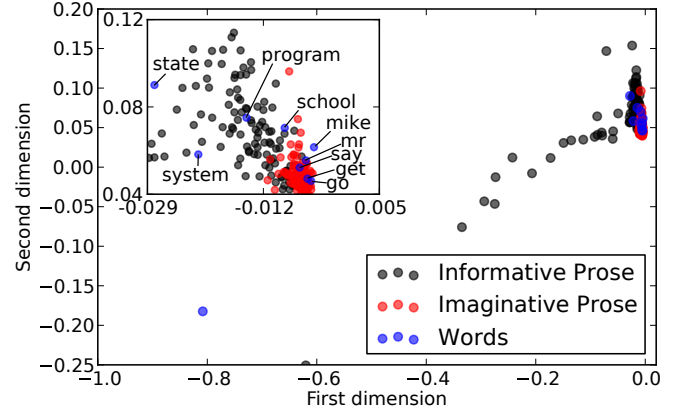


FIG. 4. Latent semantic analysis performed to distinguish informative from imaginative prose. First ten more frequent words were used as features. Note that the style can be identified by measuring the proximity to specific words. While *state* and *system* characterize informative documents, *say* and *mr.* characterize imaginative texts.

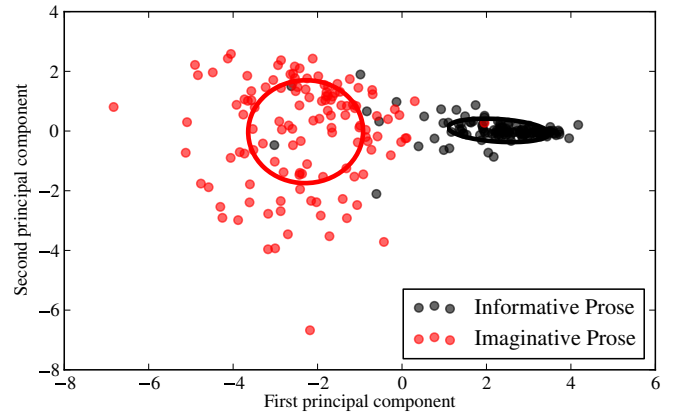


FIG. 5. Principal component analysis performed to distinguish informative from imaginative prose. The frequency of the most informative *stopwords* were used as features. Note that the style developed in informative documents is much more regular than the style observed in imaginative texts.

Considering the most informative bigrams, the accuracy rate reached 98%. A visualisation of the data provided by this set of features is shown in Fig. 6.

All in all, the results obtained with traditional classifiers demonstrate that the local topological approach is effective as our best results differs only 3% from the most efficient traditional system. This result is consistent with similar studies showing that the topology plays a relevant role in characterising complex systems, especially those conveying information [2, 12, 37, 44]. Because the proposed representation is complementary to the traditional approaches, we advocate that the combination of features

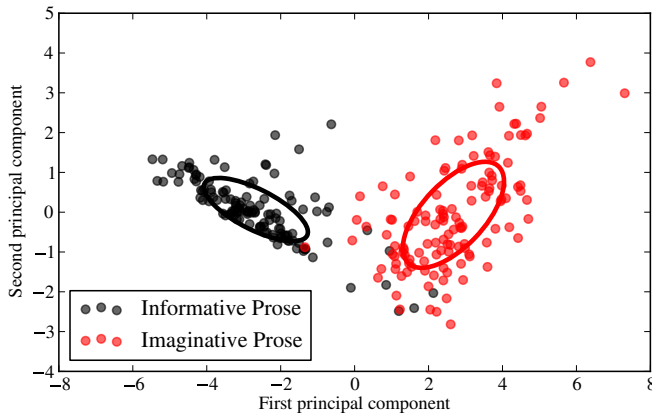


FIG. 6. Principal component analysis performed to distinguish informative from imaginative prose. The frequency of character bigrams were used as features. As it happened in Fig. 5, the variability of styles is much higher in imaginative texts.

of distinct nature (traditional and topological) could lead to the improvement of similar tasks relying on the accurate characterisation of stylistic marks.

V. CONCLUSION

In this paper, we have evaluated the ability of network measurements to identify two textual categories, which are related to informative and imaginative documents. We have extended previous models in a twofold manner. First, the local topology of nodes representing specific words was studied. We have thus emphasized particular network regions to characterize the local topology of texts. This approach differs from previous networked representations because traditional topological analyses consider with equal relevance the topological analysis of all nodes of the network. Another proposed extension is the use of novel network measurements that are able to grasp more relevant information than traditional measurements. Particularly, we have used symmetry measurements that are able to quantify the homogeneity of access to neighbours. The concept of node degree was also extended via introduction of accessibility measurements, which are able to measure the *effective* number of (accessed) neighbours.

Computational simulations revealed that the proposed extensions are able to improve the efficiency of classification tasks. The best improvement in performance when comparing the traditional model and the proposed method occurred with the *KNN* classifier. An improvement of 23% was observed, thus confirming the efficiency of the proposed methodology. A systematic analysis of feature relevance revealed that among the most infor-

mative attributes are the symmetry and accessibility indexes applied to particular nodes. These results confirm the complementary role played by these measurements in characterizing text networks, since they do not correlate with traditional natural language processing methods. Owing to the generality of the proposed representation and characterization, we believe that it could be extended to a myriad of related applications where the quantification of style is relevant for text categorization. As further works, we intend to combine network methods and traditional statistical methods to improve the performance of the classification. We expect, in this case, that the interwoven combination of methodologies will be able to overcome the limitations of each technique.

VI. ACKNOWLEDGMENTS

HFA thanks CAPES for financial support. LdFC is grateful to CNPq (Brazil) (grant no. 307333/2013-2), FAPESP (grant no. 11/50761-2), and NAP-PRP-USP for sponsorship. DRA acknowledges financial support from São Paulo Research Foundation (FAPESP) (grant no. 14/20830-0).

* Corresponding author: diego@icmc.usp.br

- [1] W. J. Hutchins and H. L. Somers, *An introduction to machine translation*, Vol. 362 (Academic Press London, 1992).
- [2] D. R. Amancio, M. G. Nunes, O. N. Oliveira, and L. d. F. Costa, *Physica A: Statistical Mechanics and its Applications* **391**, 1855 (2012).
- [3] C. C. Yang and F. L. Wang, *Journal of the American Society for Information Science and Technology* **59**, 887 (2008).
- [4] K. Patil and P. Brazdil, *Int. J. Comput. Sci. Inform. Syst* **2**, 18 (2007).
- [5] D. Marcu, *The theory and practice of discourse parsing and summarization* (MIT press, 2000).
- [6] Y.-L. Chen, Y.-H. Liu, and W.-L. Ho, *Journal of the American Society for Information Science and Technology* **64**, 280 (2013).
- [7] A. Singhal, *IEEE Data Eng. Bull.* **24**, 35 (2001).
- [8] K. Krippendorff, *Content analysis: An introduction to its methodology* (Sage, 2012).
- [9] F. Sebastiani, *ACM computing surveys (CSUR)* **34**, 1 (2002).
- [10] Y. Aphinyanaphongs, L. D. Fu, Z. Li, E. R. Peskin, E. Efsthadiadis, C. F. Aliferis, and A. Statnikov, *Journal of the Association for Information Science and Technology* **65**, 1964 (2014).
- [11] D. R. Amancio, O. N. Oliveira Jr, and L. d. F. Costa, *Physica A: Statistical Mechanics and its Applications* **391**, 4406 (2012).
- [12] D. R. Amancio, O. N. Oliveira Jr, and L. da Fontoura Costa, *New Journal of Physics* **14**, 043029 (2012).

- [13] D. Fetterly, M. Manasse, and M. Najork, in *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004* (ACM, 2004) pp. 1–6.
- [14] T. A. Meyer and B. Whateley, in *CEAS* (Citeseer, 2004).
- [15] R. Navigli, *ACM Computing Surveys (CSUR)* **41**, 10 (2009).
- [16] D. R. Amancio, O. N. Oliveira Jr, and L. d. F. Costa, *EPL (Europhysics Letters)* **98**, 18002 (2012).
- [17] T. C. Silva and D. R. Amancio, *EPL (Europhysics Letters)* **98**, 58001 (2012).
- [18] Q. Miao, Q. Li, and D. Zeng, *Journal of the american society for information science and technology* **61**, 2288 (2010).
- [19] S. A. Golder and M. W. Macy, *Science* **333**, 1878 (2011).
- [20] L. d. F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiquiera, M. P. Viana, and L. E. Correa Rocha, *Advances in Physics* **60**, 329 (2011).
- [21] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).
- [22] R. F. i Cancho and R. V. Solé, *Proceedings of the Royal Society of London B: Biological Sciences* **268**, 2261 (2001).
- [23] L. Antiquiera, M. d. G. V. Nunes, O. Oliveira Jr, and L. d. F. Costa, *Physica A: Statistical Mechanics and its Applications* **373**, 811 (2007).
- [24] D. R. Amancio, S. M. Aluisio, O. N. Oliveira Jr, and L. d. F. Costa, *EPL (Europhysics Letters)* **100**, 58002 (2012).
- [25] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr, and L. d. F. Costa, *PloS one* **8**, e67310 (2013).
- [26] R. F. i Cancho, R. V. Solé, and R. Köhler, *Physical Review E* **69**, 051915 (2004).
- [27] E. Stamatatos, *Journal of the American Society for information Science and Technology* **60**, 538 (2009).
- [28] R. Mihalcea and D. Radev, *Graph-based natural language processing and information retrieval* (Cambridge University Press, 2011).
- [29] G. A. Miller, *Communications of the ACM* **38**, 39 (1995).
- [30] L. d. F. Costa, *International Journal of Modern Physics C* **15**, 371 (2004).
- [31] J. Cong and H. Liu, *Physics of life reviews* **11**, 598 (2014).
- [32] B. B. Greene and G. M. Rubin, *Automatic grammatical tagging of English* (Department of Linguistics, Brown University, 1971).
- [33] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, *Advances in Physics* **56**, 167 (2007).
- [34] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, *Physics reports* **424**, 175 (2006).
- [35] D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).
- [36] T. Opsahl and P. Panzarasa, *Social networks* **31**, 155 (2009).
- [37] D. R. Amancio, E. G. Altmann, O. N. Oliveira Jr, and L. da Fontoura Costa, *New Journal of Physics* **13**, 123024 (2011).
- [38] L. Freeman, *Sociometry* **40**, 35 (1977).
- [39] M. E. J. Newman, *Social networks* **27**, 39 (2005).
- [40] S. P. Borgatti, *Social networks* **27**, 55 (2005).
- [41] L. C. Freeman, S. P. Borgatti, and D. R. White, *Social networks* **13**, 141 (1991).
- [42] A. E. Motter and Y.-C. Lai, *Physical Review E* (2002).
- [43] B. Mirzasoleiman, M. Babaei, M. Jalili, and M. Safari, *Physical Review E* **84**, 046114 (2011).
- [44] D. R. Amancio, *Journal of Statistical Mechanics: Theory and Experiment* **2015**, P03005 (2015).
- [45] M. Girvan and M. E. Newman, *Proceedings of the national academy of sciences* **99**, 7821 (2002).
- [46] L. C. Freeman, *Social networks* **1**, 215 (1979).
- [47] E. Estrada, *The structure of complex networks: theory and applications* (Oxford University Press, 2011).
- [48] P. Bonacich, *American journal of sociology*, 1170 (1987).
- [49] A. N. Langville and C. D. Meyer, *Google's PageRank and beyond: the science of search engine rankings* (Princeton University Press, 2011).
- [50] R. Mihalcea, P. Tarau, and E. Figa, in *Proceedings of the 20th international conference on Computational Linguistics* (Association for Computational Linguistics, 2004) p. 1126.
- [51] B. A. N. Travençolo and L. d. F. Costa, *Physics Letters A* **373**, 89 (2008).
- [52] C. E. Shannon, *ACM SIGMOBILE Mobile Computing and Communications Review* **5**, 3 (2001).
- [53] G. F. de Arruda, A. L. Barbieri, P. M. Rodríguez, F. A. Rodrigues, Y. Moreno, and L. d. F. Costa, *Physical Review E* **90**, 032812 (2014).
- [54] J. R. Finnerty, *International Journal of Developmental Biology* **47**, 523 (2003).
- [55] H. Longuet-Higgins, *Molecular Physics* **6**, 445 (1963).
- [56] P. Holme, *Physical Review E* **74**, 036107 (2006).
- [57] L. Rossi, A. Torsello, E. R. Hancock, and R. C. Wilson, *Physical Review E* **88**, 032806 (2013).
- [58] F. N. Silva, C. H. Comin, T. K. Peron, F. A. Rodrigues, C. Ye, R. C. Wilson, E. Hancock, and L. d. F. Costa, *arXiv preprint arXiv:1407.0224* (2014).
- [59] D. R. Amancio, F. N. Silva, and L. d. F. Costa, *EPL (Europhysics Letters)* **110**, 68001 (2015).
- [60] M. E. J. Newman, *Proceedings of the National Academy of Sciences* **103**, 8577 (2006).
- [61] T. M. Mitchell, Burr Ridge, IL: McGraw Hill **45** (1997).
- [62] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, 2006).
- [63] J. C. Krzysztof, W. S. Roman, and L. Kurgan, “Data mining: A knowledge discovery approach,” (2007).
- [64] C. M. Bishop, *Pattern recognition and machine learning* (springer, 2006).
- [65] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees* (CRC press, 1984).
- [66] R. J. Lewis, in *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California* (2000) pp. 1–14.
- [67] L. Getoor and B. Taskar, *Introduction to statistical relational learning* (MIT press, 2007).
- [68] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing* (MIT press, 1999).
- [69] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. d. F. Costa, *PloS one* **9**, e94137 (2014).
- [70] I. Guyon, *Feature extraction: foundations and applications*, Vol. 207 (Springer Science & Business Media, 2006).
- [71] H. B. Mann and D. R. Whitney, *The annals of mathematical statistics*, 50 (1947).
- [72] W. N. Francis and H. Kucera, Brown University (1979).

- [73] C. H. Comin, F. N. Silva, and L. d. F. Costa, EPL (Europhysics Letters) **110**, 68002 (2015).
- [74] S. T. Dumais, Annual review of information science and technology **38**, 188 (2004).